EVALUATION OF THE ATR-MATRIX SPEECH TRANSLATION SYSTEM WITH A PAIR COMPARISON METHOD BETWEEN THE SYSTEM AND HUMANS

Fumiaki Sugaya\(^{1}\), Toshiyuki Takezawa\(^{1}\), Akio Yokoo\(^{2}\), Yoshinori Sagisaka\(^{1}\), Seiichi Yamamoto\(^{1}\)

\(^{1}\)ATR Spoken Language Translation Research Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Email: fumiaki.sugaya@slt.atr.co.jp

\(^{2}\)NTT Communication Science Laboratories, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

The main goal of the present paper is to propose a new scheme for the overall evaluation of a speech translation system that supports the design of target application systems and determines their performance. Evaluations are conducted on the Japanese-to-English ATR-MATRIX speech translation system, which was developed at ATR Interpreting Telecommunications Research Laboratories. In the proposed scheme, the system is compared with native Japanese taking the Test of English for International Communication (TOEIC) for speech translation capability. A regression analysis using evaluation results shows that the speech translation capability of ATR-MATRIX matches Japanese scoring around 500 on the TOEIC.

1. INTRODUCTION

ATR Interpreting Telecommunications Research Laboratories developed the ATR-MATRIX speech translation system \([1]\), which translates both ways between English and Japanese. At ATR-SLT, we have been carrying out overall evaluations through dialog tests and analyses of this system \([2]\). In the basic hotel reservation task/domain, we have shown the effectiveness of the system. Dialog tests are effective for evaluating the system. However, they do have demerits too, e.g., a lot of labor is required like test control, transcription, and tagging. It is difficult to enlarge evaluation target domains/tasks in the same way. Furthermore, measures are necessary to support the design of target application systems to meet performance expectations.

Machine translation systems have been evaluated with A, B, C and D ranks \([3]\). This rank evaluation method is useful for making relative comparisons of the systems in time series and between several schemes. The method, however, has a demerit of not having a direct relationship with the objective performance of real application target systems. Tomita \([4]\) proposed a new scheme using the Test of English as a Foreign Language (TOEFL) to evaluate the quality of translated text as a whole. Evaluation results with this scheme can support the design of target application systems and determine their performance. The scheme, however, cannot be applied to apply to present speech translation systems, because its task/domain is limited.

We propose a new method that is applicable to speech translation systems with a limited task/domain capability. In this method, both the system and humans with variable translation capability answer questions on translations of test utterances taken from the target task/domain. The answers are compared by native evaluators. The winning rate tendency of the system’s translations matches that of humans’ in some points. Regression analysis clarifies the precise points. In section 2, the new method is explained. In section 3, the result for the language translation subsystem (TDMT) is presented. In section 4, results using speech recognition (SPREC) are shown. The effect of recognition errors on the language translation subsystem is also discussed. An accuracy analysis for the proposed method is made in section 5. Section 6 shows effects of evaluators and summarizes evaluation results. A conclusion is given in section 7.

2. TRANSLATION PAIR COMPARISON METHOD

Figure 1 shows a diagram of the proposed translation pair comparison method in the case of Japanese to English translation. The examinees are asked to listen to Japanese text and provide English translations on a paper. Japanese text is announced twice in a minute, there is a pause in-between. To measure the English capability of the Japanese native, the TOEIC score is used. The examinees are presented with an official TOEIC score certificate which shows that they have officially taken the test within six months. A questionnaire is given to them and the results show us that the answer time is moderate for the examinees.

The test text is the SLTA1 test set, which consists of 330 utterances in 23 conversations from a bilingual travel conversation database \([5]\). The SLTA1 test set is open for both speech recognition and language translation. The answers written on the pieces of paper are typed. In the proposed method, the typed translation results by the examinees and the outputs of the system are merged to make evaluation sheets, and are compared by native Americans. The evaluation sheets show two translation results: the results of the examinee and the system in random order to eliminate discrimination by the native Americans and Japanese test text. The native Americans are asked to follow the procedure in Fig. 2. The four ranks are the same as those used in \([3]\). The meanings of ranks A, B, C, and D are as follows: (A) Perfect: no problems in both information and grammar; (B) Fair: easy-to-understand with some unimportant information missing or flawed grammar; (C) Acceptable: broken but understandable...
Choose A, B, C, and D rank

No

EVEN

Same rank? Yes

Consider naturalness

No

Select better result

Yes

Same?

No

with effort; (D) Nonsense: important information has been translated incorrectly.

3. EVALUATION OF LANGUAGE TRANSLATION SUBSYSTEM

3.1 Evaluation Results of Language Translation Subsystem

Figure 3 shows a comparison between the language translation subsystem (TDMT) and the examinees. The input for TDMT included accurate transcriptions. The total number of examinees was thirty with five people at every hundred TOEIC points between 300s and 800s. Although we advertised for five examinees at each hundred TOEIC points, only a couple of score holders hit 895. The horizontal axis in Figure 3 shows the TOEIC score and each bar against a TOEIC score shows an evaluation result. The bar consists of three parts. From the horizontal line, we have the number of TDMT won utterances, the number of even (non-winner) utterances which indicates no difference between the results of TDMT and humans, and the number of examinee won utterances. These three numbers sum to 330 (the total number of test utterances). English native speakers able to understand Japanese judged the evaluation sheets. Figure 3 shows that the TDMT system wins around TOEIC scores of 300 and 400. To examinees, in contrast, win at scores around 800. The capability balanced area is around a score of 600 and 700. To precisely get the balanced point, we use regression analysis.

To prepare the regression analysis, the number of even utterances is divided and put into the number of TDMT won utterances and examinee won utterances. The dotted line in Figure 3 shows this modified number of TDMT won utterances. The straight line shows the regression line. The capability balanced point between the TDMT subsystem and the examinees is one where the regression line crosses half the number of all test utterances (330/2=165). In Figure 3, the point is 707.6. Consequently, the translation capability of the language translation system equals that of the examinees at around a score of 700 points on the TOEIC. We call this point the system’s TOEIC score.

3.2 Feature of Language Translation Subsystem

The number of TDMT won utterances is larger than that of examinee won utterances for lower TOEIC scores. In this area, the system dominates the match. The dominance rate (R) is defined by

\[
R = \frac{N_{\text{TDMT}} + \frac{N_{\text{EVEN}}}{2}}{N_{\text{HUMAN}} + \frac{N_{\text{EVEN}}}{2}}
\]  

(1)
where \( N_{TDMT} \) is the number of TDMT won utterances, \( N_{HUMAN} \) is the number of examinee won utterances, and \( N_{EVEN} \) is the number of even utterances. A rate of more than one indicates that the TDMT’s capability is superior to the examinees’ . In contrast, a rate of less than one indicates that the TDMT’s capability is inferior to the examinees’ . If the rate equals one, this indicates that the TDMT’s capability matches that of the human. Figure 4 shows the dominance rate according to the average word entropy.

In Figure 4, the dominance rate is averaged for every hundred TOEIC scores. Around lower values of the TOEIC scores, the rate is large, which means that the TDMT system dominates in capability, while around higher values of the entropy, the rate is less than 1, which indicates that the human capability is superior to that of the TDMT system. Through a corpus analysis on our bilingual travel conversation database [5], the percentage of utterances with an entropy of 4 or less is so large, i.e., 62.5%, that TDMT effectively works for the travel conversation.

### 4. EVALUATION RESULTS FOR LANGUAGE TRANSLATION WITH SPEECH RECOGNITION

Figures 5 and 6 show evaluation results of the TDMT language translation capability using speech recognition. All of the characteristics in the figures are almost similar to Figs.3 and 4. However, the system’s TOEIC scores drop into 548.0, which is lower by 150, compared with the case of accurate transcriptions for the TDMT input. This degradation is due to the speech recognizer’s performance.

Here, we define the dominance degradation rate as a ratio of the dominance rate of TDMT with a speech recognizer, and that of TDMT with accurate transcriptions, as \( R(SPREC+TDMT)/R(TDMT) \). Figure 7 shows that the speech recognition rate (WA: Word Accuracy) drastically falls with an increase in the entropy. The dominance degradation rate also shows a similar decrease. The correlation between the speech recognition rate and the dominance degradation rate is very high, i.e., 0.91. The dominance degradation rate is approximated as 1-(1-WA)*2.15.

### 5. ERRORS IN THE SYSTEM’S TOEIC SCORE

The number of modified winning utterances for the system (\( X_i \)) and TOEIC scores for the examinees (\( Y_i \)) are assumed to satisfy the population regression equation:

\[
Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad (i = 1,2,\ldots,n)
\]

where \( \beta_1, \beta_2 \) are population regression coefficients. The error term (\( \varepsilon \)) is assumed to satisfy the following condition:

\[
\begin{align*}
(a) \quad E(\varepsilon_i) &= 0 \\
(b) \quad V(\varepsilon_i) &= \sigma^2, \quad i = 1,2,\ldots,n \\
(c) \quad Cov(\varepsilon_i, \varepsilon_j) &= E(\varepsilon_i \varepsilon_j) = 0 \quad \text{if } i \neq j \\
(d) \quad \sum \varepsilon_i &= 0
\end{align*}
\]
Under the above condition, the standard deviation of the system’s TOEIC score is calculated by

$$
\sigma_T = \left| \frac{\sigma}{\sqrt{n}} \right| \sqrt{1 + \frac{C_0 - \bar{X}^2}{\sum (X_i - \bar{X})^2}}
$$

(4)

where \( n \) is the number of examinees, \( C_0 \) is the system’s TOEIC score, and \( \bar{X} \) is the average of the examinees’ TOEIC scores. Equation (4) indicates that the minimum error is given when the system’s TOEIC score equals the average of the examinees’ TOEIC scores.

By using a t-distribution, the confidence interval of the system’s TOEIC score with confidence coefficient \( 1-\alpha \) is given by

$$
\left[ C_0 - t_{1,\alpha/2} \sqrt{\frac{\sigma^2}{n}} - 2\kappa_{2T}, C_0 + t_{1,\alpha/2} \sqrt{\frac{\sigma^2}{n}} - 2\kappa_{2T} \right]
$$

(5)

### 6. ERRORS AMONG EVALUATORS

For the evaluation of the language translation subsystem with the speech recognizer, another evaluator also compared the evaluation sheets. The results had the same tendency. A regression analysis showed that the system’s TOEIC score was 544.1, while the score was 548.1 for the other evaluator. The difference was as small as 3.7. In Table 1, numerical data calculated from the evaluation results and half of the confidence interval are summarized. Half of the confidence interval was 49.45, 56.73, or 45.92 for three cases when the confidence coefficient was 0.99.

### 7. CONCLUSION

We proposed a translation pair comparison method for a speech translation system. This method is applicable to wider tasks/domains without additional labors like dialog tests. We evaluated the ATR-MATRIX system. The results showed that the system’s capability equals that of native Japanese at around a score of 548 points of the TOEIC. According to public information on the TOEIC, the average TOEIC score of university students in Japan is 568 points. Even considering the confidence interval, ATR-MATRIX is nearly approaching the average university students’ speech translation capability, which is achieved by tremendous costs even in the limited task/domain involved.

The system’s performance in language translation, speech recognition, and its effects on language translation are strongly related to the entropy. This entropy is a measure of information, and should not be directly related to the task/domain. Accordingly, the performance dependency on the entropy can be expected to be valid in other tasks/domains. The results in this paper are taken and analyzed in a limited domain/task, but the entropy dependency of the performance can hopefully be used effectively to select new tasks/domains.

### Table 1: Summary of evaluation results

<table>
<thead>
<tr>
<th></th>
<th>TDMT</th>
<th>SPREC+TDMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 )</td>
<td>307.60</td>
<td>265.64</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.20</td>
<td>-0.18</td>
</tr>
<tr>
<td>system’s TOEIC score</td>
<td>707.56</td>
<td>548.05</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>17.21</td>
<td>19.63</td>
</tr>
<tr>
<td>Average TOEIC score of examinees</td>
<td>606.05</td>
<td></td>
</tr>
<tr>
<td>variance of TOEIC scores</td>
<td>32056.92</td>
<td></td>
</tr>
<tr>
<td>the number of examinees</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>( \sigma_T )</td>
<td>17.90</td>
<td>20.53</td>
</tr>
<tr>
<td>confidence interval/2</td>
<td>49.45</td>
<td>56.73</td>
</tr>
</tbody>
</table>

### REFERENCES